The 16th International Command and Control Research and Technology Symposium (ICCRTS)
21-23 June 2011, Quebec City, Canada

"Collective C2 in Multinational Civil-Military Operations"

Paper No. 035

Supporting C2 Research and Evaluation: An Infrastructure and its Potential Impact

Primary Topic 6: Experimentation, Metrics, and Analysis

Authors: James B. Law, Ph.D. and Marion G. Ceruti, Ph.D.

Organization: Space and Naval Warfare Systems Center Pacific (SSC Pacific)

Address: 53560 Hull Street, San Diego, CA 92152-5001, USA, (619) 553-2449

Email: jim.law@navy.mil, marion.ceruti@navy.mil

James B. Law, Ph.D., Point of Contact
(619) 553 2449

Filename: Law SupportingC2Research 16thICCRTS 2011.paper35.doc

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Supporting C2 Research and Evaluation: An Infrastructure and its Potential Impact	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Space and Naval Warfare (SPAWAR) System Center (SSC) Pacific,53560 Hull Street,San Diego,CA,92152-5001**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The lifecycle management of Command and Control and Command, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) technologies depends on test data for research, development, testing, evaluation, maintenance, and sustainment. Creating the infrastructure necessary to support these activities is difficult and expensive. Progress has been slow and empirical data on the cost and effectiveness of C4ISR technologies remain poorly documented. Many data sets are becoming obsolete because they do not conform to new data formats that enable integration, fusion, flexibility, and reasoning. These data sets will require significant transformational effort to remain useful. To address these problems the design and construction of an infrastructure is in progress to support controlled experimentation with emerging C4ISR technologies. This paper describes the challenges to researchers and evaluators of C4ISR systems. The paper features a survey of publicly available data sources, characterizations of data suitability, and any required transformation. A repository is in progress to address these challenges. The paper concludes with a discussion of the potential impact of this infrastructure on C4ISR research and technology.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**25**	

Supporting C^2 Research and Evaluation: An Infrastructure and its Potential Impact

James B. Law, Ph.D. and Marion G. Ceruti, Ph.D.
Space and Naval Warfare Systems Center Pacific (SSC Pacific)
53560 Hull Street, San Diego, CA 92152-5001, USA
jim.law@navy.mil, marion.ceruti@navy.mil

Abstract *— The lifecycle management of Command and Control and Command, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C^4ISR) technologies depends on test data for research, development, testing, evaluation, maintenance, and sustainment. Creating the infrastructure necessary to support these activities is difficult and expensive. Progress has been slow and empirical data on the cost and effectiveness of C^4ISR technologies remain poorly documented. Many data sets are becoming obsolete because they do not conform to new data formats that enable integration, fusion, flexibility, and reasoning. These data sets will require significant transformational effort to remain useful. To address these problems the design and construction of an infrastructure is in progress to support controlled experimentation with emerging C^4ISR technologies. This paper describes the challenges to researchers and evaluators of C^4ISR systems. The paper features a survey of publicly available data sources, characterizations of data suitability, and any required transformation. A repository is in progress to address these challenges. The paper concludes with a discussion of the potential impact of this infrastructure on C^4ISR research and technology.*

Keywords —data aggregation, data fusion, data integration, infrastructure, integration tools, lifecycle management, repository, test-data sets

1. Introduction

Readily available, high-quality, test data would benefit nearly every aspect of Command and Control (C^2) and Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C^4ISR) system life cycles. Yet, such test data are scarce in both research and practice. A subjective assessment is that time spent searching for, organizing, storing, and reusing useful test data has a significant impact on research and proof-of-concept testing. Lack of suitable test data represents a bottleneck for developing new systems, evaluating and fielding emerging systems, and maintaining currently fielded systems. Finally, test data requirements change during the system lifecycle. Data formats and data content are changed frequently to accommodate changes in system functionality and new or modified interfaces.

In open literature, empirical studies of C^2 and C^4ISR systems are rare [27], [21], [35], [24]. Controlled experiments are even rarer [5]. We believe that this situation is exacerbated by poor data set availability and by problems associated with the effort of adapting data sets to new experiments. Controlled experiments provide the best evidence that our systems meet design and performance criteria. A common body of test data that can be adapted more easily to varying experimentation will likely provide research and practitioners with greater power to evaluate algorithms, architectures, and implementations. Therefore, a common repository of unclassified and open sourcedata sets and supporting infrastructure is proposed here.

The paper is organized as follows. Section 2 covers the specific problems the proposed repository is expected to address. Section 3 summarizes a survey of open terrorism data sources, and other related data sources. Section 4 describes the infrastructure that the proposed repository will entail. This section includes an example of how suitable test-data sets can be created from openly available data sets based on initial experimentation. Section 4 also outlines the documentation and supporting tools that will be required. Section 5 presents a discussion of concerns regarding repository use. Section 6 concludes the paper with a summary of the expected benefits of a repository.

2. Challenges for Research and Evaluation

Researchers in software testing found five primary challenges when establishing a repository of software artifacts [15]. The authors' current research and other experiences corroborate this result. The challenges are replicating findings, supporting aggregation of findings, reducing costs, obtaining representative operational profiles, and isolating the effects of individual factors. Each of these challenges is discussed below as they relate to C^4ISR systems.

2.1 Replicability of findings

Performance investigations and comparative evaluation of systems may require replicating the findings of earlier experiments, either in subsequent experiments, by other researchers, or both. Several factors can lead to failure to replicate finding. These factors include changes in the system under test; changes, ambiguities, and inconsistencies in the test data; and the experiment design or the execution environment. System test and execution environments are rarely similar across experiments or laboratories. A reusable body of well documented, well-understood test data, along with standard tools could reduce the differences significantly. Inability to replicate findings will also tend to exacerbate the problems described below.

2.2 Supporting aggregation of findings

Under certain circumstances, experimental results from different contexts may be aggregated, or combined, thus yielding additional power to the findings of the experiments. In many cases, the context of the experiments is not known well enough to support aggregation. Opportunities for aggregation are highly correlated with the replicability of findings (Section 2.1), since it is likely that the easily replicable experiments are also well documented and well understood.

2.3 Reducing costs

Creation, integration, fusion, and maintenance of data sets in general is challenging [4], time consuming, and therefore, expensive. This is especially true for test-data sets, which must be reliable as future life-cycle decisions will be made on the basis of software testing. Simple reuse of data sets can alleviate this problem somewhat. However, it does not solve the problems of data-set changes and the problems of finding new data sets to keep pace with the demand of changing system requirements. A community repository of data sets, documentation, experiment designs, and other infrastructure could reduce costs significantly, enable more testing, and promote more critical examination of systems.

2.4 Obtaining representative operational profiles

The number of potential execution paths in software can be exponential in the size of the program or software system. Since most C^4ISR systems are relatively large, the number of execution paths is effectively infinite. Effective module and system testing relies on sampling from test cases or operational profiles to gain reasonable assurance that the software system is reliable [31]. Test suites and operational profiles that are not representative may endanger research and development milestones and possibly the end users of the system.

Test data sets and operational profiles typically are developed by isolated groups for their own use. Since the development effort is significant, the results tend to meet minimum requirements. A repository such as that described here could address the issues of both of quantity and quality.

2.5 Isolating the effects of individual factors

Independent variables must be isolated in a controlled experiment in order to discover causality relationships. Given the expense of controlled experiments, it could be said that uncovering causality relationships is why we tolerate the expense of performing experiments. Ad-hoc and opportunistically developed data sets and operational profiles frequently lack the breadth or depth to exercise enough of the independent variables' space to make reasonable determinations of causality.

3. Survey of Data Sources

Like software, data sets are developed generally for a specific purpose, which limits their reuse. (See, for example, [12].) A significant part of the present study included a survey to ascertain the availability of data sets and resources of interest to the C^4ISR community. The first part of this survey focused on terrorism and counter-terrorism data, whereas the second part included data sets that would be useful for other purposes.

The survey yielded the following sources of terrorism-event data online. Most of the data sets are small and use simple text formats. The organizations that control the data sets usually determine who can access them. Anyone who needs the data sets can contact the organizations to request access.

University of Arizona, Dark Web Terrorism Research Project [13]. The University of Arizona Artificial Intelligence Lab systematically scraped web sites that hosted discussions of Islamic ideology and religion. The contents are hosted in the Dark Web Forum Portal, which contains messages from 28 forums, for a total of approximately 13 million messages. The messages are primarily in text format with some documents and video. The research group has performed a variety of analyses of the links, content, web metrics, sentiment, authorship, and video. The portal host has limited visualization tools. Access to the raw data also is cumbersome.

As is the case with the majority of data sources described below, the Arizona AI Lab Dark Web project provides an aggregation of raw-data sets, generally in text format. Data aggregation is the lowest level in the data-integration and data-fusion hierarchy [5]. Whereas many of these data sets are large and often have significant extent, they lack the explicit identification of complex interrelationships that results from data-integration and data-fusion processing. In almost all cases, the reason for this is that a substantial, costly, and labor-intensive effort is required to discover and maintain these complex interrelationships, especially at the semantic level of data integration [8], [10]. However, higher-level data integration and fusion processes, such as semantic integration, are required to produce the best functionality in C^4ISR systems. (See, for example, [11] and [9].)

Haverford Database of Terrorist Acts. The Global Terrorism Resource Database [30] maintains a collection of statements issued by Al-Qaeda leaders and a variety of other terrorism documents in plain text and pdf formats. The documents are primarily speech transcriptions, policy statements, recruitment materials, and/or propaganda. The reliability of the content of these documents is open to question due to factors such as the potential for deception, misdirection, factional disagreements, and changes in leadership. Therefore, the usefulness of these documents for some purposes, such as indications of intent and prediction, is questionable. Haverford also maintains an extensive list of additional online data resources [30], including other terrorism data collections, some of which are discussed below.

Center for Defense and International Security Studies (CDISS), Database of Terrorist Incidents, 1940 – 1999 [3]. CDISS provides a summary of terrorist incidents by decade from 1940 through 1999. Each incident is summarized in one or two sentences. The information provided on each incident is extremely limited.

French Database of Terrorist Acts (Base De Donnes Sur Les Actes Terroristes [17]. This data set is a record of the details of terrorist attacks since 1965 against France, or French interests, and in global areas of French interest. The agency maintaining this data set requires access arrangements in the form of a user ID and password. The data may not be downloaded. The website was unavailable at the time of this writing.

NIST Message Understanding Conference (MUC) Archives [34]. The Message-Understanding Conferences (MUC) were organized as an exercise in extracting meaning from test messages. The conferences supplied a prepared data set in text format and called for participants to perform specific tasks such as entity recognition, co-reference recognition, and other information extraction tasks. The participants were evaluated by a panel that had access to objective answers. The conference series originally was organized and supported by Science Applications International Corporation (SAIC) [39]. The proceedings of MUC-3 through MUC-6 were published by Morgan Kaufman Publishers [14]. They were out of print at the time of this writing. The NIST website archives the MUC-7 Proceedings from 1999, several earlier data sets, as well as definitions and example tasks. The data sets for the last two conferences (MUC-6 and MUC-7) are available for a fee from the Linguistic Data Consortium [45], which is discussed below.

RAND Database of Worldwide Terrorism Incidents [11]. RAND maintains and extensive database of covering the last 30 years of worldwide terrorism activity. Details on each incident are limited. The RAND website requires registration.

SEMVAST - Scientific Evaluation Methods for Visual Analytics Science and Technology [40, 41]. SEMVAST is a series of yearly experiments from 2006 through 2010. The participants in the experiment receive a synthetic data set. The experimenters ask participants to make determinations that depend on the information domains, which have included text, cell phone network data, wiki edits, simulated employee badge and corporate network traffic, and genetic sequences.

National Consortium for the Study of Terrorism and Responses to Terrorism (START) [46]. START maintains two databases of terrorism incidents. The Global Terrorism Database contains information on 87,000 worldwide terrorist events from 1970 through 2008. The information provided on each incident is limited, but much more extensive than events in the

CDISS [3] database discussed above. The START consortium also maintains a database of Terrorist Organization Profiles, which the consortium wants to integrate with their Global Terrorism Database. Although both databases are available with a query interface, no apparent method is provided for downloading the results of queries or for aggregating and analyzing these results on the website. However, interested parties may request data from the consortium via the website.

Terrorism and Preparedness Data Resource Center [44]. Terrorism incidents that involve the United States and its allies are a growing concern of US government agencies. Researchers at the START consortium [46] host, maintain, and manage this resource at the Inter-University Consortium for Political and Social Research, ICPSR [43].

National Institutes of Justice (NIJ) – Terrorism Databases for Analysis [33]. The NIJ maintains a database of US-domestic incidents for analysis. Whereas this database provides an additional catalog of incidents, the information lacks detail.

Terrorism in Western Europe: Events Data (TWEED) [16]. The TWEED data source contains a summary of terrorism incidents in Western Europe. As with other data sets described in this section, TWEED containing summaries of incidents and the data sets are limited.

Joint Threat Anticipation Center [42]. JTAC is a collaborative project of the University of Chicago Center for International Studies and Argonne National Laboratory that developed an integrated database on the worldwide history of terrorist groups and a database of worldwide terrorist incidents.

Worldwide Incident Tracking System (WITS) [32]. The following is a direct quote from the introductory material: "The Worldwide Incidents Tracking System (WITS) contains data available to the public from the National Counterterrorism Center (NCTC), presenting it in a navigable and searchable user interface. With the interface users can run reports and queries to retrieve analytic information related to attacks worldwide, filter the results, produce charts, and view results on maps." The National Counterterrorism Center (NCTC) manages this resource.

Ali Baba Data Set [22]. This is a synthesized data set created by the Department of Defense. The data contain information on nine hidden groups in a moderate size body of messages. It has been used in a number of initial studies in government, academia, and industry. The free availability of this data set is not known at the time of this writing.

Counter-Terror Social Network Analysis and Intent Recognition (CT-SNAIR). CT-SNAIR [48] is a project of MIT Lincoln Laboratory. The project uses a variety of social-network-analysis and machine-learning techniques to analyze collected data. The initial data set used by the project is from the September 2004 bombing of the Australian embassy in Jakarta. Other data sets that they used were from a prior research project and the Ali Baba data set [22]. The free availability of this data set is not known at the time of this writing.

World-Trade Center Event Sequence Data. The Department of Sociology and the Institute for Mathematical Behavioral Sciences at the University of California, Irvine have created an event sequence data set from transcripts of the radio communications of first responders at the World Trade Center terrorism attack on September 11, 2001. The format of the data is an R-archive (http://r-project.org). A description of the data is available in [23]. This data source is nearly unique in providing fine-grained data from actual terrorist events.

The following data sets may prove useful in future experiments, although they are not specifically related to terrorism.

Inter-University Consortium for Political and Social Research, ICPSR [43]. ICPSR hosts a large number of small data sets that were used in social and political research studies. The primary deficiency of these data sets for the purpose of the present study is their small size and isolated coverage.

The book, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms* [1], contains several small but interesting data sets.

ClueWeb09 Dataset [2]. From the introduction, The "ClueWeb09" data set was created by the Language Technologies Institute at Carnegie Mellon University to support research on information retrieval and related human-language technologies. The data set consists of 1 billion web pages written in ten languages, collected in January and February 2009. The data set is used by several tracks of the TREC conference." The data sets are available on four 1.5 TB hard disks for a fee of approximately $750, which covers the maintenance expenses, cost of the hard disks, shipping, and handling.

Terrorism Data Sets	Sources	Size	Formats
Dark Web Terrorism Research Project [13]	Web crawls	Large	Text, video
Haverford Database of Terrorist Acts [30]	News, propaganda	Medium	Text, pdf
(CDISS), Database of Terrorist Incidents, 1940 – 1999 [3]	CDISS	Large	Text
French Database of Terrorist Acts [17]	French agencies	Meduim	Text
Message Understanding Conference (MUC) Archives [34]	Synthetic	Small	Text
RAND Database of Worldwide Terrorism Incidents [11]	RAND	Medium	Text
Scientific Evaluation Methods for Visual Analytics Science and Technology (SEMVAST) [40, 41]	Synthetic	Medium	Formatted text
National Consortium for the Study of Terrorism and Responses to Terrorism (START) [46]	START	Large	Databases, text
Terrorism and Preparedness Data Resource Center [44]	START	Medium	Text
NIJ Terrorism Databases for Analysis [33]	NIJ	Small	Databases
Terrorism in Western Europe: Events Data (TWEED) [16]	TWEED	Medium	Database
Joint Threat Anticipation Center (JTAC) [42]	JTAC	Medium	Database
Worldwide Incident Tracking System (WITS) [32]	NCTC	Medium	Database
Ali Baba Data Set [22]	Synthetic	Small	Text
Counter-Terror Social Network Analysis and Intent Recognition (CT-SNAIR) [48]	Govt agencies	Medium	Text
World-Trade Center Event Sequence Data [23]	UC Irvine [23]	Small	R-archive

Table 1. Summary of terrorism related data sets.

Linguistic Data Consortium (LDC) [45]. The LDC provides a selection of linguistic data files in SGML format. The data files are copyrighted and the LDC requires a fee for use.

Stanford large network dataset collection [25]. This collection contains several large and interesting graph and network data sets. The collection maintainer also distributes a C++ software package for graph and social-network analysis named SNAP.

MemeTracker data [26]. The MemeTracker data sets are a record of phrases and hyperlinks extracted from blog posts and news articles between August 2008 and April 2009. There information is divided into two data sets, one organized by phrase clusters and the second containing the raw phrase data for each source posting or new article.

UC Irvine Machine Learning Repository [18]. The archive currently maintains 189 different data sets. Sixteen of these sets contain low-dimensionality time-series data. Particularly interesting sets include the *CalIt2 Building People Counts Data Set* [19] and the *Dodgers Loop Sensor Data Set* [20], which could be

Non-Terrorism Data Sets	Sources	Size	Formats
Inter-University Consortium for Political and Social Research, ICPSR [43]	Research studies	Many small sets	Various
Modeling the Internet and the Web: Probabilistic Methods and Algorithms [1]	Research studies	Small	Formatted text
ClueWeb09 Dataset [2]	Web crawls	Large	Text
Linguistic Data Consortium (LDC) [45]	LDC	Small	Text
Stanford large network dataset collection	Various	Large	Formatted text
MemeTracker data	News, blogs	Large	Formatted text
UC Irvine Machine Learning Repository	Research studies	Small	Formatted text

Table 2. Summary of non-terrorism related data sets.

used to test techniques for finding anomalies in observed traffic or activity.

We have summarized the Terrorism-related and Non-terrorism related data sets in Table 1 and Table 2.

In many cases multiple data sets described above can be combined to yield a larger data set that would be useful for limited testing. For example, an interesting and potentially important exercise would be to search for available links in the CIDSS and START data sources to detailed information on each of these incidents in other terrorism data collections. Long-term correlations can be identified from analysis of the aggregate of multiple historical data sources, e.g. CIDSS and START, in such a longitudinal study. These correlations probably would not be evident from the examination of single-source material produced during much more limited timeframes. The aggregate and fusion of these data might yield a more comprehensive picture of terrorist trends, strategy, and the evolution of terrorist tactics.

4. Required Infrastructure

Personnel involved in C4ISR testing need improved, aggregated, integrated, and fused data sets to address the challenges discussed above in Section 2 and in [4], as well as the limitations of the data sources described in Section 3. This section describes an example of aggregating, integrating and fusing data sets to produce a larger, more complex data set. The construction of an infrastructure for a repository of data sets and associated resources is in progress at SSC Pacific. The authors divide the proposed repository infrastructure into two parts. The first part consists of the data sets, whereas the second part consists of the documentation and supporting tools for using, sharing, and extending the repository. Subsections 4.1 and 4.2 cover the significant problems and strategies associated with each part respectively.

4.1 Data Sets.

The data sets described in Section 3 are formatted predominantly as plain text. An informal canvassing of authors' colleagues indicates that the expected future data-set format is Resource Description Format (RDF) [51]. Whereas other alternatives have some merit, RDF has many advantages over plain text. The primary advantage of RDF is the standardization of parsers and supporting tools. Both RDF and plain-text formats can represent rich data. However, each plain-text format will require a custom parser. In contrast, well-developed RDF parsers are available for most widely used computer languages. Therefore, RDF is most likely to be the preferred data format for our repository.

In the remainder of this subsection we show how to aggregate, integrate, and transform raw test-data sets into a form that is more useful for research, development and testing of C^4ISR systems. The simple example to illustrate this point is based on two machine-learning data sets. This section also contains

a discussion of criteria for selecting new test-data sets, creating synthetic test data sets, and evaluating test-data sets.

When our research group began working with applications that require RDF-formatted data, the format of most available data was plain text. To save time, we transformed several small files into RDF format using the simplest structure that reflected the data. For the example we present below, we used two data files from the UCI Machine Learning Repository [18]. The first file contained data from observations of human traffic in and out of the CalIt2 building on the University of California, Irvine (UCI) campus in Irvine, California [19]. The observations were collected over 15 weeks for a total of 10,080 observations. The RDF that resulted from the transformation is depicted symbolically on the left side of Figure 1. The RDF blank nodes, depicted as ovals and labeled "_a," each represent a 30-minute interval of time. Each of the two red circles connected to the ovals by the arrowheads represents the number of people who went into and out of the CalIt2 building, respectively.

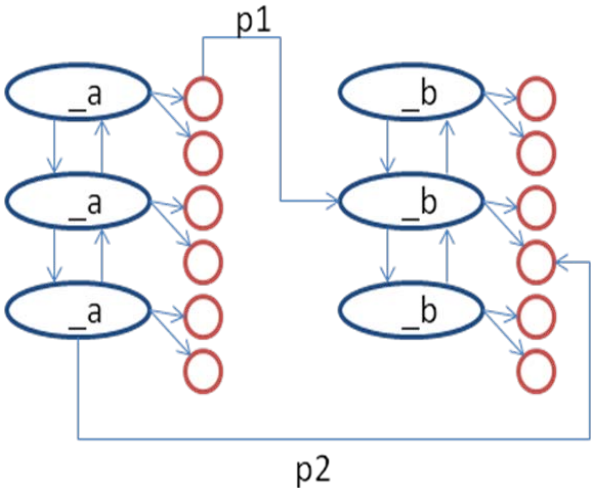


Figure 1. Schematic depiction of simulated correlated RDF tracking based on observed data sets

The second data file we used was the Dodger's Loop Sensor Data Set [20], also from the UCI Machine Learning Repository. The Dodger's Loop counts the number of cars passing over the Glendale on-ramp of the 101 North Freeway in Los Angeles, CA near Dodger's Stadium. The number of cars passing this point in subsequent 5-minute time intervals was collected over 25 weeks for a total of 50,400 observations. The RDF that resulted from our transformation of this file is shown on the right of the drawing in Figure 1. RDFLib [38] was used for the Python programming language to read the CalIt2 and the Dodger's Loop Sensor data files, and to perform the transformations.

These two initial RDF files sufficed for testing our preliminary prototypes and learning about RDF formats. However, these data sets lack the depth and complexity of data elements used in common C4ISR environments. Unfortunately, we could not find more complex data sets that were openly available. Therefore, we had to manufacture more complex RDF data sets by randomly linking elements in the two previously prepared files to simulate a correlated data set, as shown by the predicates "p1" and "p2" in Figure 1. Obviously, the resulting correlated files are logically and physically unrealistic from a human standpoint. However, they work well for testing supporting scripts and utilities, and for exploring possible RDF structures for new data-set development. Since these are now, in effect, synthetic data sets, we can alter the type and density of predicate linkages at will. From this initial combined set we eventually created a family of test data sets of varying complexity and depth. Future transformations of this simulated "data mashup" and additional data will provide a deeper understanding of data complexity in testing. The programming effort required to realize these synthetic data sets is not trivial. However, we have observed the following subjective benefits:

- For applications which lack usable data, ad-hoc data sets can be created relatively quickly.
- Small existing data sets can be selectively enlarged.
- Prospective algorithms and systems can be easily tested for sensitivity to node set size and edge density.
- These synthetic data sets can be readily shared with collaborators.

These exercises have also helped us gain experience with RDF, understand transformation processes better, and explore potential RDF representations for larger real-world C^4ISR data sets. We intend to apply the experience gained and the supporting infrastructure created to transforming large internally developed data sets that have obsolete formats.

4.2 Documentation and Supporting Tools

Two keys to the reuse of software artifacts are adherence to standards [12], [11] and complete well-written documentation [36]. This is still generally true for reusing and integrating data sets. Good metadata

documentation [8], [10], [4], including pedigree, includes but is not limited to the following:

- A complete specification of formats (including standard formats),
- Date-time group (e.g. day, hour, minute, sec.) of data-element or data-set collection,
- Latitude and longitude or other location specification of data collection,
- Data pedigree, e.g. the origin of the data (e.g. sensor ID) [7], [47]
- The means (e.g. sensors, observations, etc.) that were used to collect the data sets,
- The providence or processes involved in producing or delivering the data [52].
- The algorithms that were used to integrate and fuse the data,
- Standard metrics such as data-set size.

Subsequent users, such as a commander's staff in a command center, need pedigree metadata to evaluate and use the data set, and to reduce uncertainty in decision making [6]. Supporting tools include scripts or specifications required for storing, retrieving, and managing the data. Another important form of documentation is previous experimental designs. Good experimental design is not always obvious to inexperienced researchers and evaluators [50]. Consideration of proven experimental designs can avoid having to repeat experiments based on bad designs that can invalidate the results. Good experimental design also can aid the replication of the valid results as discussed above in Section 2.1.

Methods for using, sharing, and extending data sets and their supporting infrastructure vary greatly and usually are created ad hoc. Thus, they need to be re-created each time a new group uses the data set for purposes that differ from the original use. A repository would support reuse of infrastructure, such as scripts. In a net-centric, service-oriented environment, it is possible to share processes and procedures for use with a variety of data sets [49], [47].

5. Discussion

Our first concern is that reusing data sets poses threats to both internal and external validity which must be well understood when designing and conducting experiments [29]. Internal validity is an indication of how well the experiment controls the dependent variables and, therefore, how strong the causal relationships in the experiment can be trusted. External validity concerns the extent to which the results of the experiment can be generalized to other experimental subjects. One of the functions of data set documentation and previous studies should be to highlight these hazards for future users.

Our second concern is that extending and integrating the repository is likely to happen over time, by the users of the repository, if a sufficient level of use occurs. The exact nature and requirements of the infrastructure are likely to be defined in practice rather than known beforehand. We can expect the content and use of the repository to change as algorithms, technologies, and system architectures change.

6. Conclusions, Expected Benefits, and Future Research

The proposed repository has significant development expense, however we expect it will result in a larger number of empirical studies, and potentially higher quality empirical studies. Although we currently lack empirical measures of cost reduction or improved efficacy, we expect to gather and report this evidence in future. The authors feel that the existence of a data set repository such as we propose will likely encourage such empirical studies. We further think it is reasonable to expect this repository to benefit all phases of the C4ISR system lifecycle.

Future infrastructure development work includes integrating XML and RDF validation support and visualization services. We will also be adding our current data set creation scripts to the initial repository and soliciting feedback from test users.

Our future research will continue developing RDF format data sets using openly available data. We have also located several existing and open RDF data sets. We will be integrating these exiting RDF sets with ones we have created. Our intent is to create several RDF data sets for community use that reasonably reflect the size, complexity, and loosely correlated events with ground truth that are needed for continuing research, development, and fielding of C^4ISR systems. We encourage researchers and practitioners with similar interests to contact the authors.

Acknowledgements

The authors thank the U.S. Office of Naval Research for financial support of this project. The authors also acknowledge the help of numerous colleagues at SSC Pacific and at the Naval Research Laboratories at Washington, DC and Stennis for reviews, suggestions, and discussion. This paper is the work of U.S.

Government employees performed in the course of employment and no copyright subsists therein. It is approved for public release with an unlimited distribution.

References

[1] P. Baldi, P. Frasconi, and P. Smyth, *Data sets for the Book Modeling the Internet and the Web*, http://ibook.ics.uci.edu/datasets.html

[2] Carnegie Mellon University Language Technologies Institute, *The ClueWeb09 Dataset*, http://boston.lti.cs.cmu.edu/Data/clueweb09/

[3] CDISS, *Database of Terrorist Incidents, 1940 – 1999*, http://www.cdiss.org/pages/Programmes/Revolutionary_Warfare_Counter_Insurgency/Publications.asp

[4] M.G. Ceruti, "Data Management Challenges and Development for Military Information Systems," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. Vol. 15 No. 5, pp. 1059-1068, September/October 2003. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA455063&Location=U2&doc=GetTRDoc.pdf http://portal.acm.org/citation.cfm?id=942752

[5] M.G. Ceruti, "An Expanded Review of Information-System Terminology." *Proceedings of the AFCEA Federal Database Colloquium '99*, pp. 173-191, September 1999, San Diego CA. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA371047&Location=U2&doc=GetTRDoc.pdf

[6] M.G. Ceruti, A. Ashfelter, R. Brooks, G. Chen, S. Das, G. Raven, M. Sudit, and E. Wright, "Pedigree Information for Enhanced Situation and Threat Assessment," *Proceedings of the 9th IEEE International Conference on Information Fusion (FUSION 2006)* July 2006, Firenze Italy.

[7] M.Ceruti and J. Kaina, "Enhancing Dependability of the Battlefield Single Integrated Picture Through Metrics for Modeling and Simulation of Time-Critical Scenarios." *Proceedings of the Ninth International IEEE Workshop on Object-oriented Real-time Dependable Systems*, (WORDS 2003F), pp. 69-76, October 2003, Anacapri Italy.

[8] M.G. Ceruti and M.N. Kamel, "Semantic Heterogeneity in Database and Data Dictionary Integration for Command and Control Systems," *Proceedings of the DOD Database Colloquium '94*, pp. 65-89, August 1994, San Diego CA. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA289180&Location=U2&doc=GetTRDoc.pdf

[9] M.G. Ceruti, M.N. Kamel and B.M. Thuraisingham, "Object-Oriented Technology for Integrating Distributed Heterogeneous Database Systems," *Proceedings of the 12th DOD Database Colloquium '95*, pp. 79–98, August 1995, San Diego CA.

[10] M.G. Ceruti and M.N. Kamel, "Preprocessing and Integration of Data from Multiple Sources for Knowledge Discovery," *International Journal on Artificial Intelligence Tools, (IJAIT)*, Vol. **8**, No. 2, pp. 152-177, June 1999. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA364414&Location=U2&doc=GetTRDoc.pdf

[11] M.G. Ceruti, S.D. Rotter, K. Timmerman and J. Ross, "Operations Support System (OSS) Integrated Database (IDB) Design and Development: Software Reuse Lessons Learned," *Proceedings of the AFCEA Database Colloquium '92*, August 1992, San Diego CA. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA262856&Location=U2&doc=GetTRDoc.pdf

[12] M.G. Ceruti and S.D. Rotter, "Software Reuse Key Element of Navy Data Base Structure," *Signal*, Vol. 48, No. 1, pp. 55-56, 1993.

[13] H. Chen and C. Yang, eds. *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*, Springer, New York, NY. 2008. http://ai.arizona.edu/research/terror/

[14] DARPA, *Message Understanding: Evaluation and Conference: Proceedings of the 3rd-6th DARPA Workshops***,** Morgan Kaufman Publishers, 1996.

[15] H. Do, S. Elbaum, and G. Rothermel, "Supporting Controlled Experimentation with Testing Techniques: An Infrastructure and its Potential Impact," *Empirical Software Engineering*, Vol. 10 No. 4, pp. 405-435, 2005. http://sir.unl.edu

[16] J. O. Engene, *Terrorism in Western Europe: Events Data (TWEED)*, http://folk.uib.no/sspje/tweed.htm

[17] Fondation pour Recherche Stratégique, *Base De Donnes Sur Les Actes Terroristes*, Paris, France, https://bdt.frstrategie.org/

[18] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Science, 2010. http://archive.ics.uci.edu/ml/

[19] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Science, 2010.

http://archive.ics.uci.edu/ml/datasets/CalIt2+Building+People+Counts

[20] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Science, 2010. http://archive.ics.uci.edu/ml/datasets/Dodgers+Loop+Sensor

[21] J.T. Freeman, G.E. Campbell, and G. Hildebrand, "Measuring the impact of advanced technologies and reorganization on human performance in a combat information center, *Proceedings of the IEA 2000/HFES 2000 Congress*, 2000.

[22] D.A. Gerdes, C. Glymour, and J. Ramsey, "Who's Calling? Deriving Organization Structure from Communication Records," *Information Warfare and Organizational Decision Making*, ed. Alexander Kott, Artech House, Boston, MA 2007. http://www.scribd.com/doc/19012131/Information-Warfare-and-Organizational-Decision Making

[23] L. Jasny, C.S. Marcum, and C.T. Butts, "*improv: A Dataset of Disaster Response Microevents*," 2009, R package version 0.05. Contact Carter Butts (buttsc@uci.edu) for availability.

[24] D. Lafond, S. Tremblay, G. Dubé, R. Rousseau, and R. Breton, "A tool for estimating the costs/benefits of teamwork in different C2 structures, *Proceedings of the 15th International Command and Control Research and Technology Symposium (ICCRTS 2010)*, Santa Monica, CA, June 2010.

[25] J. Leskovec, *The Stanford Large Network Dataset Collection*, http://snap.stanford.edu/ data/

[26] J. Leskovec, L. Backstrom and J. Kleinberg, *MemeTracker*, http://www.memetracker.org/data.html

[27] Y. N. Levchuk, K. R. Pattipati, and D. L. Kleinman, *Analytic Model Driven Organizational Design and Experimentation in Adaptive Command and Control*, Technical Report, Defense Technical Information Center, 1999. http://handle.dtic.mil/100.2/ADA440207

[28] G.M. Levchuk and K.R. Pattipati, "Design of command and control organizational structures: from years of modeling to empirical validation," *Proceedings of the 15th International Command and Control Research and Technology Symposium (ICCRTS 2010)*, Santa Monica, CA, June 2010.

[29] R.L. Mason, R.F. Gunst, and J.L. Hess, *Statistical Design and Analysis of Experiments, with Applications to Engineering and Science*, Second Edition. Wiley-Interscience, 2003.

[30] B. Mendelsohn, *Global Terrorism Resource Database* http://people.haverford.edu/bmendels/ Mendelsohn also maintains a list of terrorism data sources: http://people.haverford.edu/bmendels/terror_attacks

[31] J.D. Musa, "Operational profiles in software-reliability engineering," *IEEE Software*, Vol. 10 No. 2, pp. 14-32, March 1993.

[32] National Counterterrorism Center (NCTC), *Worldwide Incidents Tracking System (WITS)*, https://wits.nctc.gov/

[33] National Institute of Justice (NIJ), *Terrorism Databases for Analysis*, http://www.ojp.usdoj.gov/nij/topics/crime/terrorism/databases.htm

[34] National Institute of Standards and Technology (NIST), Information Access Division (IAD), *Introduction to Information Extraction*, MUC Archive Site, http://www-nlpir.nist.gov/related_projects/muc/index.html

[35] M.E. Nissen, "Enterprise Command, Control, and Design: Bridging C2 Practice and CT Research," *The International C2 Journal*, Vol. 1, No. 1, 2007. Defense Technical Information Center, http://handle.dtic.mil/100.2/ADA486841

[36] R. Pressman, *Software Engineering: A Practitioner's Approach* 7th Ed., McGraw-Hill, 2009.

[37] RAND Corporation, *Database of Worldwide Terrorism Incidents*, http://www.rand.org/nsrd/projects/terrorism-incidents/index.html

[38] RDFlib, *A Python Library for Working with RDF*, http://rdflib.net

[39] SAIC, Inc. 1710 SAIC Drive, McLean, VA 22102. http://www.saic.com/

[40] SEMVAST Project, *Scientific Evaluation Methods for Visual Analytics Science and Technology*, http://www.cs.umd.edu/hcil/semvast/

[41] SEMVAST Project, *Visual Analytics Benchmarks Repository*, http://hcil.cs.umd.edu/localphp/hcil/vast/archive/index.php

[42] University of Chicago, *Joint Threat Anticipation Center (JTAC)*, http://jtac.uchicago.edu/resourceTerror.shtml

[43] University of Michigan, Inter-University *Consortium for Political and Social Research (ICPSR)*, http://www.icpsr.umich.edu/icpsrweb/ICPSR/

[44] University of Michigan, *Terrorism and Preparedness Data Resource Center (TPDRC)*, http://www.icpsr.umich.edu/icpsrweb/TPDRC/

[45] University of Pennsylvania, *The Linguistic Data Consortium*, http://www.ldc.upenn.edu/

[46] U.S. Department of Homeland Security, *National Consortium for the Study of Terrorism and Responses to Terrorism (START)*, http://www.start.umd.edu/start/

[47] J. Waters, M. Stelmach and M. Ceruti, "Spiral Systems Implementation Methodology:

Application of the Knowledge Web and Network-Centric Best Practices," *World Science and Engineering Academy and Society Transactions on Information Science and Applications*, Vol. 2 No. 12, pp. 2088 – 2095, December 2005. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA454877&Location=U2&doc=GetTRDoc.pdf

[48] C. Weinstein, W. Campbell, B. Delaney, and G. O'Leary, *Modeling and Detection Techniques for Counter-Terror Social Network Analysis and Intent Recognition*, 2009. http://www.ll.mit.edu/mission/communications/ist/publications/081023_CampbellW-CT-SNAIR.pdf

[49] D.R. Wilcox and M.G. Ceruti, "A Structured Service-Centric Approach for the Integration of Command and Control Components," *Proceedings of the IEEE International Conference on Service Computing (SCC 2008)*, Vol. 2, pp. 5-12, July 2008, Honolulu, HI. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04578503

[50] C. Wohlin, P. Runeson, M. Host, M. Ohlsson, B. Regnell, and A.Wesslen, *Experimentation in Software Engineering: An Introduction*, Kluwer Academic Publishers, 2000.

[51] World Wide Web Consortium (W3C), Resource Description Framework (RDF). See: http://www.w3.org/RDF/

[52] Provenance XG Final Report, W3C Incubator Group, Dec 8 2010 http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

SPAWAR
Systems Center
PACIFIC

Supporting C2 Research and Evaluation: An Infrastructure and its Potential Impact

James Law, Ph.D. and Marion Ceruti, Ph.D.

Space and Naval Warfare Systems Center Pacific (SSC Pacific)

16th ICCRTS, Quebec City, Canada
21-22 June, 2011

Topic 6: Experimentation, Metrics, and Analysis

SPAWAR Systems Center PACIFIC

Why?

- The incredible, time-consuming, difficulty of getting test data.
  - Searching for data…
  - Organizing data…
  - Storing data…
  - Reusing data…
  - Maintaining data…
- Bottleneck for:
  - Development.
  - Testing.
  - Integration.
  - Fielding.
  - Maintenance.

SPAWAR Systems Center PACIFIC

C4ISR Empirical Studies & Experiments

- Four (4) empirical studies.
- One (1) controlled experiment.
- Our Thesis:
  - A common body of test data that can be used for varying experimentation will provide researchers and practitioners with greater power to evaluate
    - algorithms,
    - architectures,
    - implementations.
- Our Proposal:
  - A common repository of unrestricted, open source data sets and supporting infrastructure.

SPAWAR Systems Center PACIFIC

Challenges for Research and Evaluation

- Replicability of findings.
- Supporting aggregation of findings.
- Reducing costs.
- Obtaining representative operational profiles.
- Isolating the effects of individual factors.

Survey of Data Sources

[1] P. Baldi, P. Frasconi, and P. Smyth, *Data sets for the Book Modeling the Internet and the Web*, http://ibook.ics.uci.edu/datasets.html
[2] Carnegie Mellon University Language Technologies Institute, *The ClueWeb09 Dataset,* http://boston.lti.cs.cmu.edu/Data/clueweb09/
[3] CDISS, *Database of Terrorist Incidents, 1940 – 1999,* http://www.cdiss.org/pages/Programmes/ Revolutionary_Warfare_Counter_Insurgency/ Publications.asp
…
[13] H. Chen and C. Yang, eds. Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security, Springer, New York, NY. 2008. http://ai.arizona.edu/research/ terror/
[14] DARPA, Message Understanding: Evaluation and Conference: Proceedings of the 3rd-6th DARPA Workshops, Morgan Kaufman Publishers, 1996.
…
[16] J. O. Engene, Terrorism in Western Europe: Events Data (TWEED), http://folk.uib.no/sspje/tweed.htm
[17] Fondation pour Recherche Stratégique, Base De Donnes Sur Les Actes Terroristes, Paris, France, https://bdt.frstrategie.org/
[18] A. Frank and A. Asuncion, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Science, 2010. http://archive.ics.uci.edu/ml/
[19] A. Frank and A. Asuncion, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Science, 2010.
…
[23] L. Jasny, C.S. Marcum, and C.T. Butts, "improv: A Dataset of Disaster Response Microevents," 2009, R package version 0.05. Contact Carter Butts (buttsc@uci.edu) for availability.
…
[25] J. Leskovec, The Stanford Large Network Dataset Collection, http://snap.stanford.edu/ data/
[26] J. Leskovec, L. Backstrom and J. Kleinberg, MemeTracker, http://www.memetracker.org/data.html
…
[30] B. Mendelsohn, *Global Terrorism Resource Database* http://people.haverford.edu/bmendels/ Mendelsohn also maintains a list of terrorism data sources: http://people.haverford.edu/ bmendels/terror_attacks
...
[32] National Counterterrorism Center (NCTC), Worldwide Incidents Tracking System (WITS), https://wits.nctc.gov/
[33] National Institute of Justice (NIJ), Terrorism Databases for Analysis, http://www.ojp.usdoj.gov/nij/topics/crime/ terrorism/databases.htm
[34] National Institute of Standards and Technology (NIST), Information Access Division (IAD), Introduction to Information Extraction, MUC Archive Site, http://www-nlpir.nist.gov/ related_projects/muc/index.html
[35] M.E. Nissen, "Enterprise Command, Control, and Design: Bridging C2 Practice and CT Research," The International C2 Journal, Vol. 1, No. 1, 2007. Defense Technical Information Center, http://handle.dtic.mil/100.2/ADA486841
…
[37] RAND Corporation, Database of Worldwide Terrorism Incidents, http://www.rand.org/nsrd/ projects/terrorism-incidents/index.html
…
[40] SEMVAST Project, Scientific Evaluation Methods for Visual Analytics Science and Technology, http://www.cs.umd.edu/hcil/semvast/
[41] SEMVAST Project, Visual Analytics Benchmarks Repository, http://hcil.cs.umd.edu/localphp/hcil/ vast/archive/index.php
[42] University of Chicago, Joint Threat Anticipation Center (JTAC), http://jtac.uchicago.edu/resourceTerror.shtml
[43] University of Michigan, Inter-University Consortium for Political and Social Research (ICPSR), http://www.icpsr.umich.edu/icpsrweb/ICPSR/
[44] University of Michigan, Terrorism and Preparedness Data Resource Center (TPDRC), http://www.icpsr.umich.edu/icpsrweb/TPDRC/
[45] University of Pennsylvania, The Linguistic Data Consortium, http://www.ldc.upenn.edu/
[46] U.S. Department of Homeland Security, National Consortium for the Study of Terrorism and Responses to Terrorism (START), http://www.start.umd.edu/start/
…

SPAWAR Systems Center PACIFIC

Survey of Data Sources

Terrorism Data Sets	Sources	Size	Formats
Dark Web Terrorism Research Project [13]	Web crawls	Large	Text, video
Haverford Database of Terrorist Acts [30]	News, propaganda	Medium	Text, pdf
(CDISS), Database of Terrorist Incidents, 1940 – 1999 [3]	CDISS	Large	Text
French Database of Terrorist Acts [17]	French agencies	Meduim	Text
Message Understanding Conference (MUC) Archives [34]	Synthetic	Small	Text
RAND Database of Worldwide Terrorism Incidents [11]	RAND	Medium	Text
Scientific Evaluation Methods for Visual Analytics Science and Technology (SEMVAST) [40, 41]	Synthetic	Medium	Formatted text
National Consortium for the Study of Terrorism and Responses to Terrorism (START) [46]	START	Large	Databases, text
Terrorism and Preparedness Data Resource Center [44]	START	Medium	Text
NIJ Terrorism Databases for Analysis [33]	NIJ	Small	Databases
Terrorism in Western Europe: Events Data (TWEED) [16]	TWEED	Medium	Database
Joint Threat Anticipation Center (JTAC) [42]	JTAC	Medium	Database
Worldwide Incident Tracking System (WITS) [32]	NCTC	Medium	Database
Ali Baba Data Set [22]	Synthetic	Small	Text
Counter-Terror Social Network Analysis and Intent Recognition (CT-SNAIR) [48]	Govt agencies	Medium	Text
World-Trade Center Event Sequence Data [23]	UC Irvine [23]	Small	R-archive

SPAWAR Systems Center PACIFIC

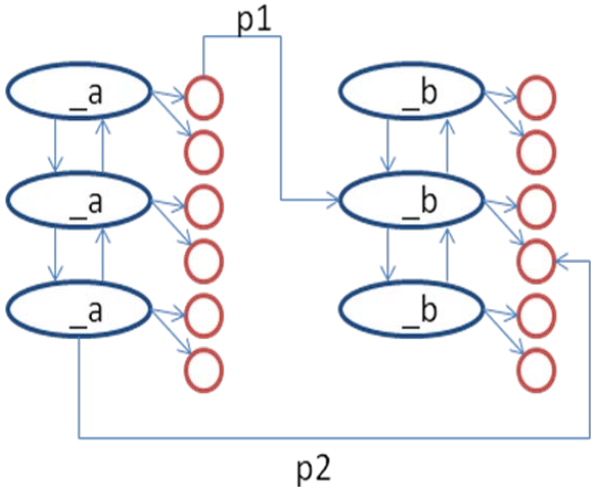
Survey of Data Sources

Non-Terrorism Data Sets	Sources	Size	Formats
Inter-University Consortium for Political and Social Research, ICPSR [43]	Research studies	Many small sets	Various
Modeling the Internet and the Web: Probabilistic Methods and Algorithms [1]	Research studies	Small	Formatted text
ClueWeb09 Dataset [2]	Web crawls	Large	Text
Linguistic Data Consortium (LDC) [45]	LDC	Small	Text
Stanford large network dataset collection	Various	Large	Formatted text
MemeTracker data	News, blogs	Large	Formatted text
UC Irvine Machine Learning Repository	Research studies	Small	Formatted text

Required Infrastructure

Our proposed repository contains:

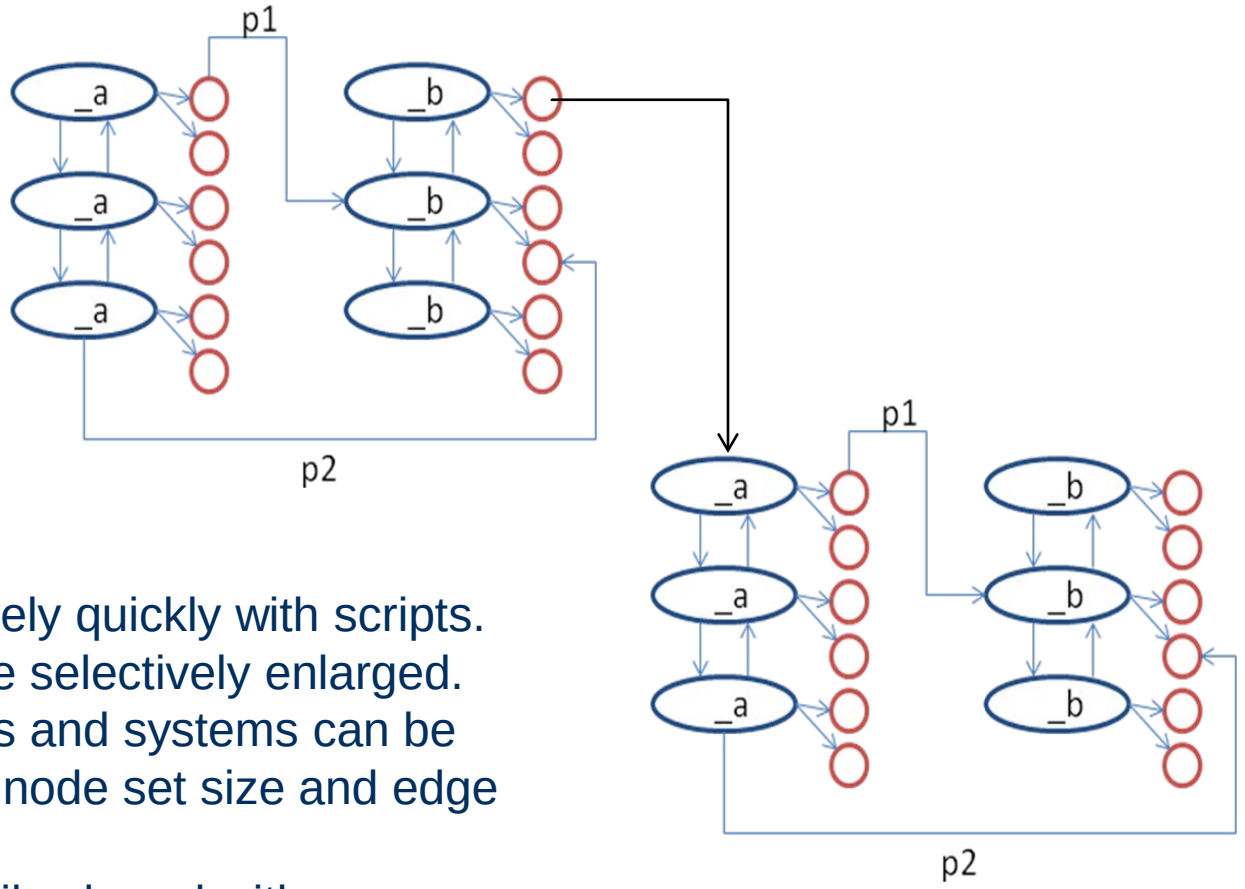
- Data Sets (RDF)



- Documentation and Supporting Tools

Data Sets

- Synthetic
- Mashups
- Real



Benefits of mashups:

- Can be created relatively quickly with scripts.
- Small data sets can be selectively enlarged.
- Prospective algorithms and systems can be tested for sensitivity to node set size and edge density.
- Data sets can be readily shared with collaborators.

SPAWAR Systems Center PACIFIC

Documentation and Supporting Tools

- A complete specification of formats (including standard formats)
- Date-time group (e.g. day, hour, minute, sec.) of data-element or data-set
- Latitude and longitude or other location specification of data collection
- Data pedigree, e.g. the origin of the data (e.g. sensor ID)
- The means (e.g. sensors, observations, etc.) used to collect the data sets
- The provenance or processes involved in producing or delivering the data
- The algorithms that were used to integrate and fuse the data
- Standard metrics such as data-set size
- Transformation scripts

SPAWAR Systems Center PACIFIC

Concerns

- Threats to validity
  - Internal - how well the experiment controls the dependent variables and, therefore, how strong the causal relationships in the experiment can be trusted.
  - External - extent to which the results of the experiment can be generalized to other experimental subjects.
- Extension and integration
  - How will existing data sets be maintained?
  - How will new data sets be integrated?
  - How will the infrastructure be adapted as systems and practice change?

SPAWAR Systems Center PACIFIC

Repository State

Current:

- Several example mashups.
  - Expect to be available for unrestricted release soon.
- Example Python scripts for transforming to RDF.

Future:

- XML and RDF validation support.
- Visualization support.
- Feedback from internal users.

jim.law@navy.mil

SPAWAR
Systems Center
PACIFIC

SSC *PACIFIC*
on Point
and at the Center of C4ISR